CMR INSTITUTE OF TECHNOLOGY

Department of Master of Computer Application

(Affiliated to Visvesvaraya Technological University)

Sri Nivasa Reddy Layout,AECSLayout,Marathalli,Bengaluru,Karnataka 560037

PROJECT REPORT

on

Lung Cancer Detection using Machine Learning

Submitted by

1) Name: Asif B
   (USN:1CR20MC017)
2) Name: Md Saddam Hussain
   (USN:1CR20MC045)

Under the Guidance of

MS. Vakula Madam

CMR INSTITUTE OF TECHNOLOGY

Department of Master of Computer Application

(Affiliated to Visvesvaraya Technological University)

Sri Nivasa Reddy Layout,AECS Layout,Marathalli,Bengaluru,Karnataka 560037

certificate

This is to certify that the report of the titled Lung Cancer Detection using Machine Learning System is a bonafide record of work done by Asif B (USN:1CR20MC017) and Md.Saddam Hussain (USN:1CR20MC017) of CMR Institute of Technology under Visvesvaraya Technological University (VTU),in partial fulfillment of the requirements of third semester MCA during the year 2021-2022.

Head of the Department

Valued by:

Date:

Name: Asif B

USN:1CR20MC045

Name: Md Saddam Hussain

USN:1CR20MC045

Exam Center:CMR Instituteof Technology

Table of contents:

ACKNOWLEDGEMENT

A project is job of great enormity and it can't be accomplished by an individual all by them Eventually, we are grateful to several individuals whose professional guidance, assistance and encouragement have made it a pleasant endeavour to undertake this project.

We endure our Humble and sincere gratitude to Ms. Gomathi T (HOD) for his great encouragement and valuable support.

We offer our sincere to our guide MS. Vakula who has always given us a constant source of inspiration and encouragement during course of our project.

Last but not the least we thank our management and lab coordinator for providing us the support to complete the project We would like to thank each and everybody who supported us throughout the long and attention consuming project.

Introduction :

Cancer is a disease in which cells in the body grow out of control. When cancer starts in the lungs, it is called lung cancer.
Lung cancer is the leading cause of cancer death and the second most diagnosed cancer in both men and women in the United States. After increasing for decades, lung cancer rates are decreasing nationally, as fewer people smoke cigarettes and as lung cancer treatments improve.
Cigarette smoking is the number one cause of lung cancer, from the moment you inhale smoke into your lungs, it starts damaging your lung tissue. The lungs can repair the damage, but continued exposure to smoke makes it increasingly difficult for the lungs to keep up the repair.

Lung cancer also can be caused by using other types of tobacco (such as pipes or cigars), breathing second-hand smoke, being exposed to substances such as asbestos or radon at home or work, and having a family history of lung cancer.

Breathing in other hazardous substances, especially over a long period of time, can also cause lung cancer. A type of lung cancer called mesothelioma is almost always caused by exposure to asbestos.

Also it was found that alcohol consumption was associated with an increased risk for lung cancer in which consuming at least 21 drinks per week. Several cases control studies have reported that alcohol consumption was associated with an increased risk for lung cancer

Objective:

Lung cancer is considered as the deadliest cancer in the world. For this reason many countries are developing the strategies for early detection of Lung Cancer. In this project the objective is to give the best result accuracy of lung cancer patients. To achieve this objective, we use KNeighbors Classifier and Decission Tree algorithm to classify the dataset and give the best accuracy of the result.

Using machine learning algorithm we predict whether the person is having Lung Cancer or not according to given dataset.

In the dataset it was shown like if the person is having Lung Cancer it is represented as 1 and if not it is represented as 0.

Software Requirement Specification

Hardware Requirements:

Hardware Tools	Minimum Requirement
Processor	I5 or above
Ram	4GB
Hard Disk	50GB

Software Requirements:

Software Tools	Minimum Requirements
Platform	Windows
Operating System	Windows 7 or above
Technology	Machine Learning – Python
Scripting Language	Python
IDE	Jupyter Note Book

Analysis and Design:

To perform KNN Algorithm and Decission Tree algorithm we use scikit-learn library. In the background of this project we used following libraries.

Numpy: The name "Numpy" stands for "Numerical Python". It is the commonly used library. It is a popular machine learning library that supports large matrices and multi-dimensional data. It consists of in-built mathematical functions for easy computations. Even libraries like TensorFlow use Numpy internally to perform several operations on tensors. Array Interface is one of the key features of this library.

Pandas: Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc.

Sklearn: It is a famous Python library to work with complex data. Scikit-learn is an open-source library that supports machine learning. It supports variously supervised and unsupervised algorithms like linear regression, classification, clustering, etc. This library works in association with Numpy and SciPy.

KNeighbors Classifier: KNN is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution. All training data used in the testing phase. This makes training faster and testing phase slower and costlier. Costly testing phase means time and memory.

Decission Tree: This is a non parametric supervised learning method used for classification and regression. The goal is to create a model that predicts value of a target variable learning simple decision rules inferred from the data features.

Matplotlib: This library is responsible for plotting numerical data. And that's why it is used in data analysis. It is also an open-source library and plots high-defined figures like pie charts, histograms, scatter plots, graphs, etc.

Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Dataset:

We have downloaded the Dataset downloaded from kaggle.com

In this project we have taken a dataset which contains 59 people's data which contains following Attributes:

Attributes of Dataset:

- Age
- Smokes
- AreaQ
- Alcohol
- Result

Smokes:

This column specifies the frequeny of smoke by the patient.

Result:

This column contains values 1 and 0 which specifies what are the chances of people getting infected with lung cancer.

IMPLEMENTATION:

Coding:

```
import warnings

warnings.filterwarnings('ignore')

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.tree import DecisionTreeClassifier

from pandas.plotting import scatter_matrix

from matplotlib import pyplot

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import confusion_matrix

from sklearn.metrics import f1_score

from sklearn.metrics import accuracy_score

from sklearn import tree


print("Dataset")

dataset=pd.read_csv(r'C:\Users\Asif Khan\Documents\pythonproject\Lung Cancer Detection\dataset.csv')

print(len(dataset))

print(dataset.head())
```

```
dataset.shape
```

```
dataset.describe()
```

```
dataset.info()
```

```
dataset.mode()
```

```
dataset.iloc[55:59,3:6]
```

```
sns.boxplot(dataset['Age']);
```

Checking whether dataset has missing values

Feature Scaling:

```
sc_x=StandardScaler()
x_train=sc_x.fit_transform(x_train)
x_test=sc_x.transform(x_test)
```

Using KNN algorithm:

```
import math
a=math.sqrt(len(y_train))
print(a)
```

Defining a model

```
classifier=KNeighborsClassifier(n_neighbors=5, p=2, metric='euclidean')
```

Fit model

```
classifier.fit(x_train, y_train)
```

predict test_set result:

```
y_pred=classifier.predict(x_test)
print(y_pred)
```

Evaluate model:

Confusion Matrix:

```
cm=confusion_matrix(y_test, y_pred)
print("confusion matrix:")
print(cm)

print('F1 Score:',(f1_score(y_test,y_pred))*100)
print("Accuracy:",(accuracy_score(y_test, y_pred))*100)
```

Using Decission Tree:

```
c = tree.DecisionTreeClassifier()
c.fit(x_train, y_train)
accu_train=np.sum(c.predict(x_train)==y_train)/float(y_train.size)
accu_test=np.sum(c.predict(x_test)==y_test)/float(y_test.size)
print('Classification accuracy on train',(accu_train)*100)
print('Classification accuracy on test',(accu_test)*100
```

Output Screenshots:

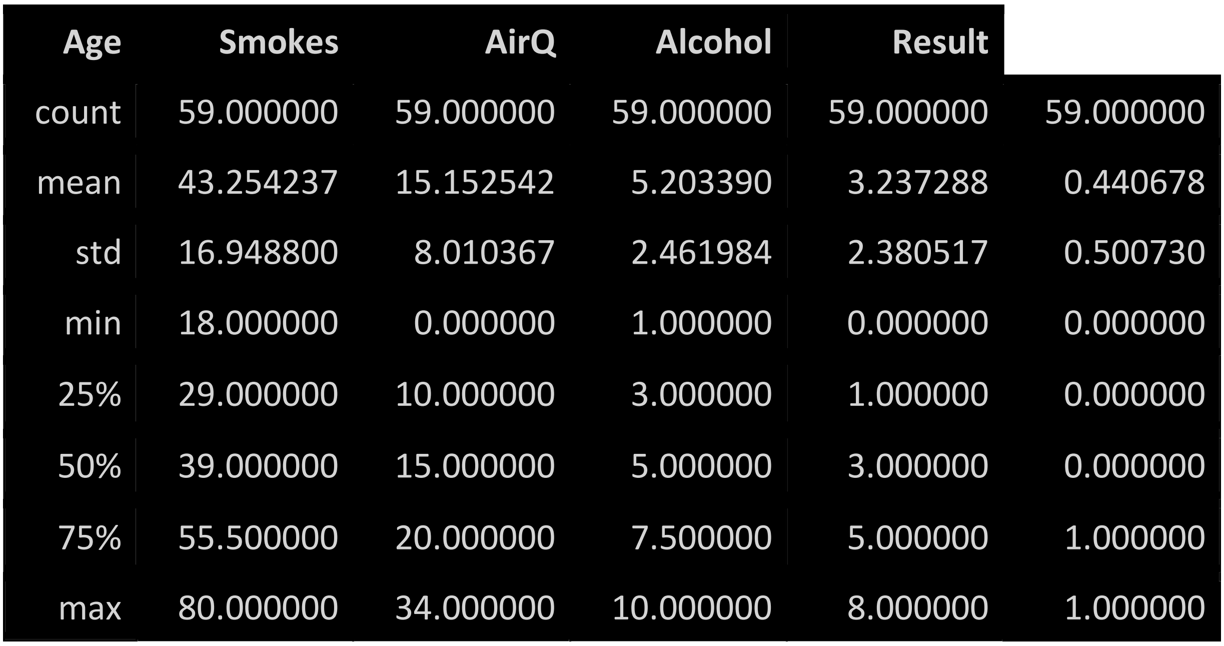
Dataset:(First 5 records of the dataset)

Name	Surname	Age	Smokes	AirQ	Alcohol	Result
John	Wick	35	3	5	4	1
jackson	Constantine	27	20	2	5	1
Camela	Anderson	30	0	5	2	0
Alex	Telles	28	0	8	1	0
Diego	Maradona	68	4	5	6	1

dataset.shape(Total number of rows and columns in the dataset)

(59,7)

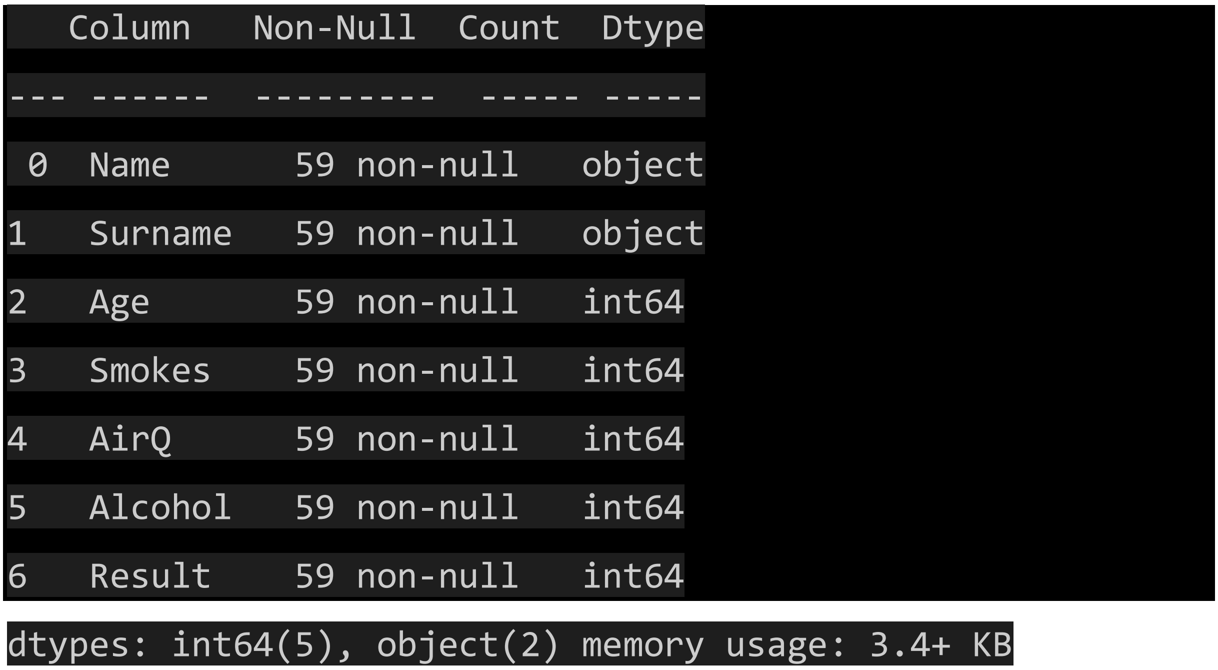
dataset.describe(): It counts the total values and finds the mean()



Age	Smokes	AirQ	Alcohol	Result		
count	59.000000	59.000000	59.000000	59.000000	59.000000	
mean	43.254237	15.152542	5.203390	3.237288	0.440678	
std	16.948800	8.010367	2.461984	2.380517	0.500730	
min	18.000000	0.000000	1.000000	0.000000	0.000000	
25%	29.000000	10.000000	3.000000	1.000000	0.000000	
50%	39.000000	15.000000	5.000000	3.000000	0.000000	
75%	55.500000	20.000000	7.500000	5.000000	1.000000	
max	80.000000	34.000000	10.000000	8.000000	1.000000	

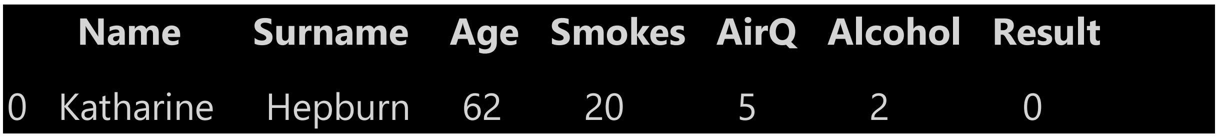
dataset.info():

(It displays information regarding the total null or not null values in the dataset)



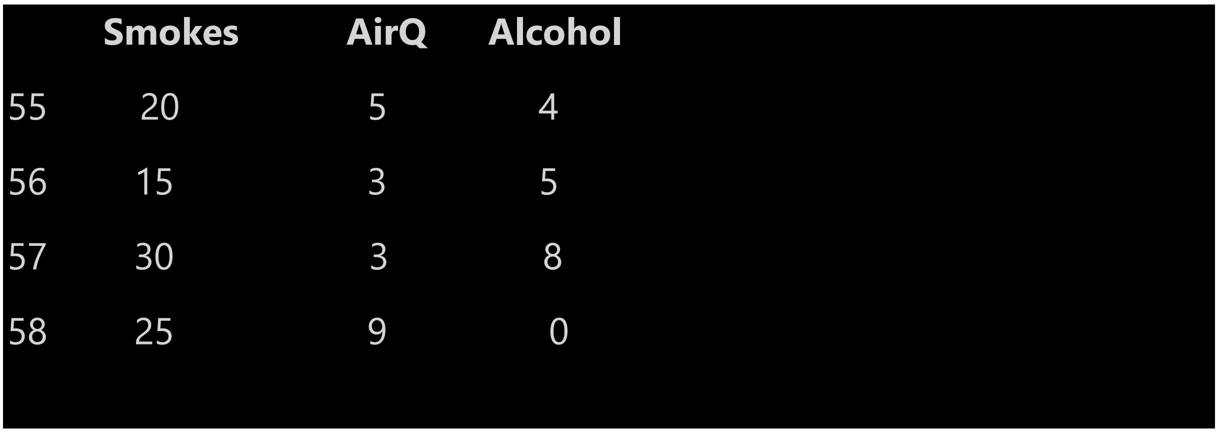
```
   Column    Non-Null  Count  Dtype
--- ------  ---------  ----- -----
 0  Name       59 non-null   object
1   Surname    59 non-null   object
2   Age        59 non-null   int64
3   Smokes     59 non-null   int64
4   AirQ       59 non-null   int64
5   Alcohol    59 non-null   int64
6   Result     59 non-null   int64
dtypes: int64(5), object(2) memory usage: 3.4+ KB
```

dataset.mode() : (Values which repeated most times in the dataset)



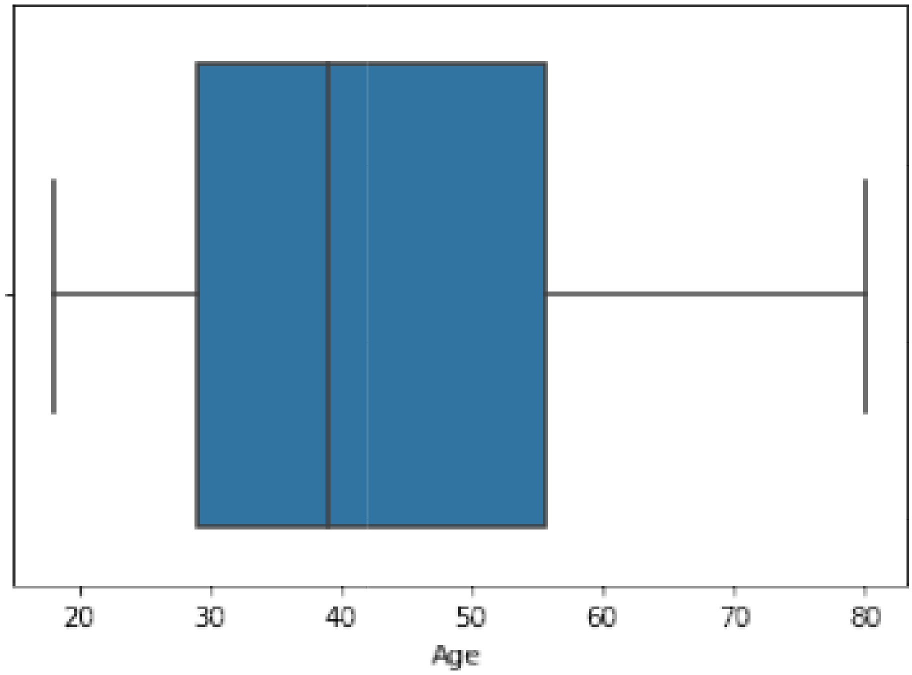
```
     Name       Surname    Age  Smokes  AirQ  Alcohol  Result
0  Katharine    Hepburn    62     20     5       2        0
```

dataset.iloc[55:59,3:6]: (extracting 3 main attributes of the dataset)



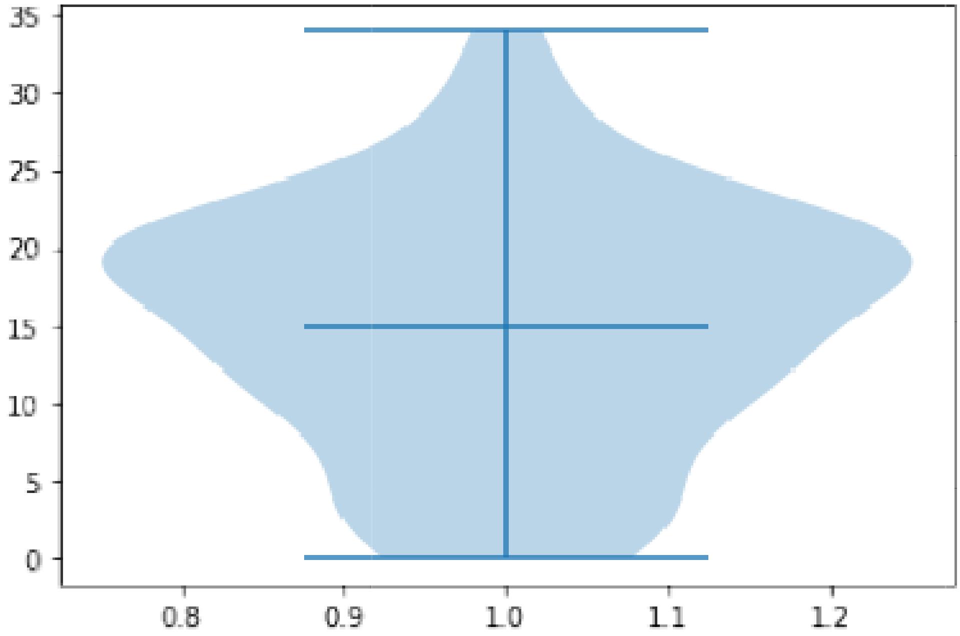
```
      Smokes       AirQ    Alcohol
55      20           5         4
56      15           3         5
57      30           3         8
58      25           9         0
```

sns.boxplot(dataset['Age']);



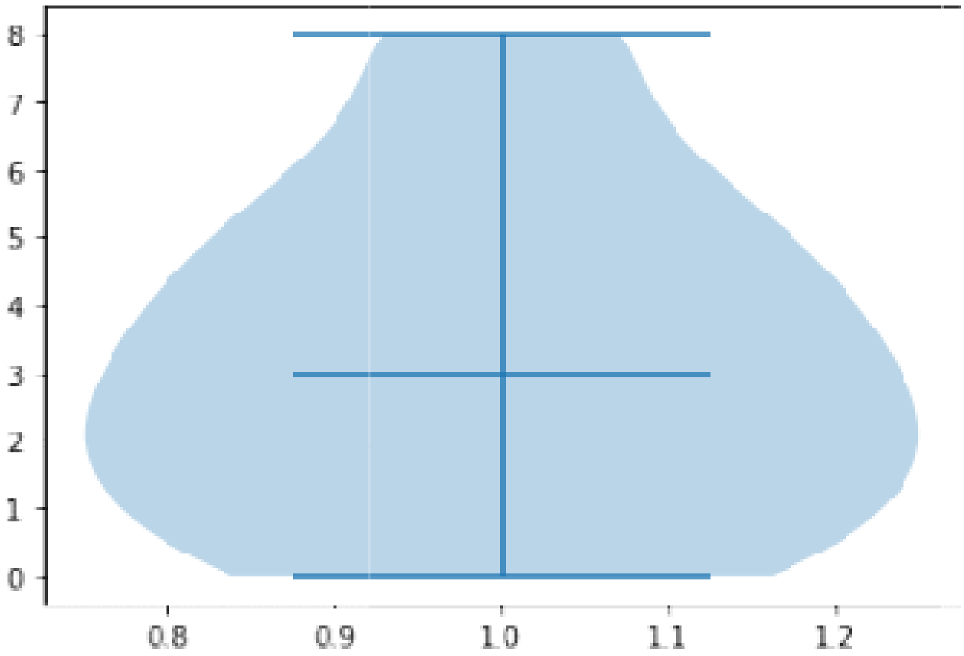
Smokes Consumption:

plt.violinplot(dataset['Smokes'],showmedians=True)



Alcohol Consumption:

```
plt.violinplot(dataset['Alcohol'],showmedians=True)
```



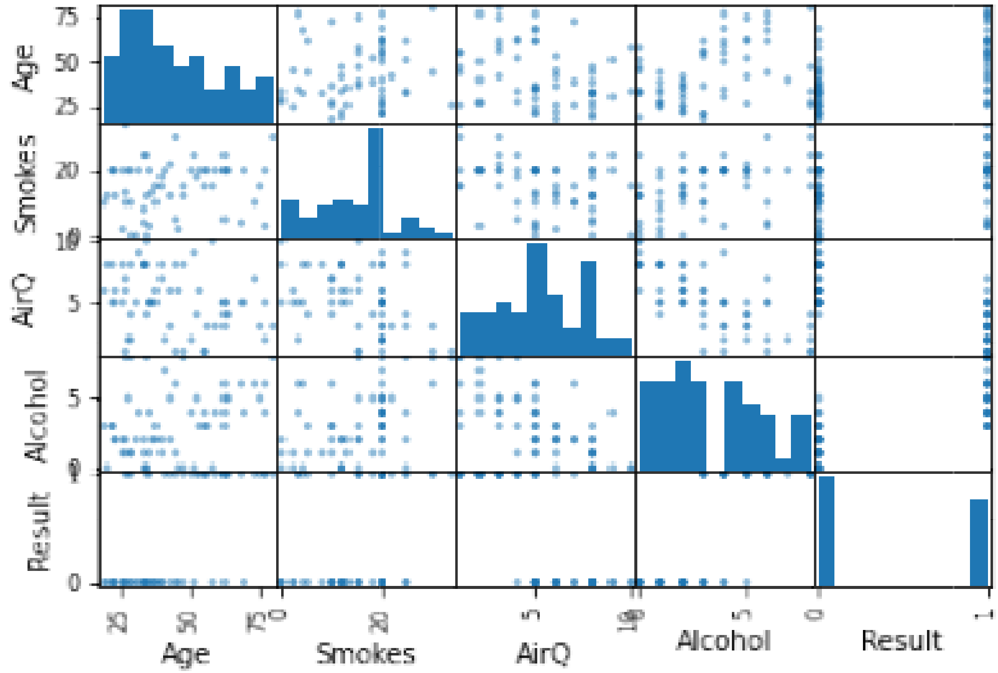
Checking whether dataset has missing values

```
dataset.isnull().sum()
```

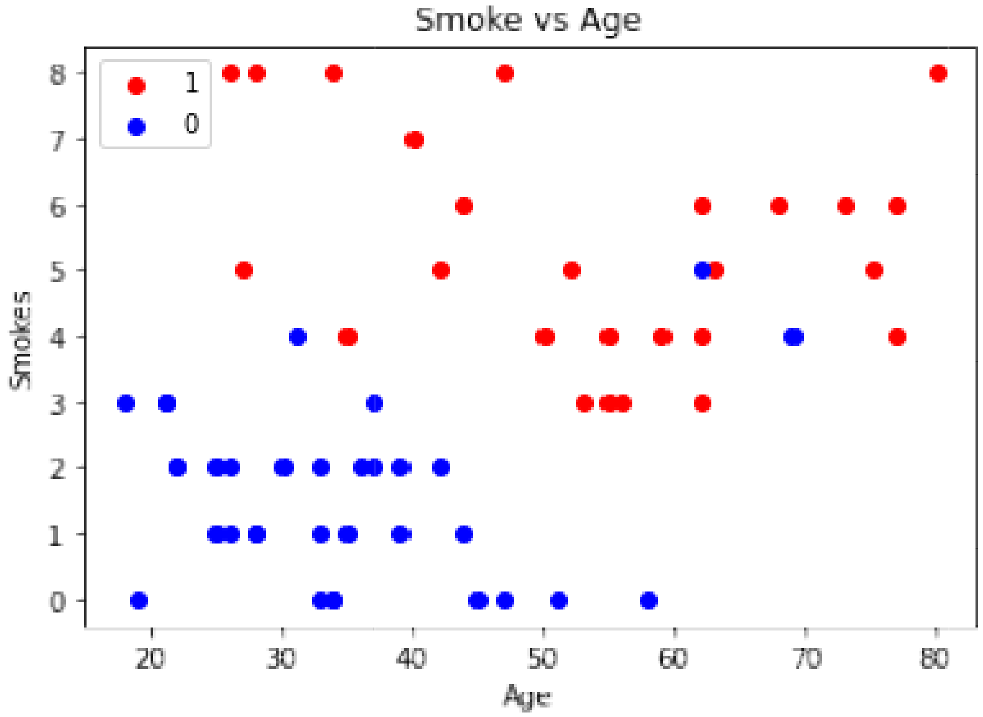
```
Name       0
Surname    0
Age        0
Smokes     0
AirQ       0
Alcohol    0
Result     0
dtype:  int64
```

scatter_matrix(dataset) for all the dataset Attributes:

pyplot.show()



Scatter plot for Smokes vs Age Comparison:



Future Scope:

The lung cancer detection system using machine learning technique is much efficient and gives the betterment results to the radiologist and assist them. This enhances with additional features for upgrading in future. On this processing system to support for radiologist to detect the affected patients as accurate result.

Machine learning is the key to enabling the Artificial Intelligence and the future of healthcare is data-driven. Big data and machine learning have a tremendous potential in the healthcare field. All these technologies are not only improving treatment and diagnosis options, they also have the potential to take control of their own health by empowering the individuals.

With the help of advanced analytics, artificial intelligence and machine learning some of the most exciting advances are coming about in healthcare. Advances in AI interfaces, personalized medicine, predictive healthcare and advances in diagnosis all come down to the application of machine learning to help the patients to access smarter healthcare techniques.

CONCLUSION:

We processed the dataset to differentiate the affected patients and its level of the growth of the Cancer by using Machine Learning System. Here it presented an approach to find best accuracy of the cancer result to assist the radiologist and for the future enhancement. Further loads ought to be directed at improving the classifying accuracy levels of the result through experiment with various alternatives.

A benchmarking of the most performing architectures on available datasets using similar metrics can help in their comparative analysis. Finally one of the current limitations is that the data and its imbalanced nature. The use of new loss functions designed to tackle the problem of unbalanced classes such as focal loss, could improve existing results, and help to achieve more efficient training. With more datasets and more balanced data, I think that better results can be achieved.

Bibiliography and Reference:

- Dataset: Lung Cancer Dataset
- Machine learning journal
- ML Algorithms(analyticsvidya.com)
- Scikit-Learn library(scikit-learn.org)